Context-Dependent Text Mining for the Retrospective Identification of High-Risk, High-Reward Research

Daniel E. Russ Ph.D.[1], Faye C. Austin Ph.D.[2], Giun Sun[1], William Lau[1], Calvin A. Johnson Ph.D.[1]
[1]Center for Information Technology, [2]Office of Portfolio Analysis and Strategic Initiatives

Abstract

In response to a Congressional mandate, the High-Risk, High-Reward (HRHR) Demonstration Project was formed to determine how NIH is addressing the challenge of funding high-risk, high-reward research. One component of this demonstration project is an assessment of the effectiveness of text mining in identifying HRHR research. In contrast with other portfolio analysis tasks, the identification of HRHR research is not dependent on extracting scientific and medical terms from controlled vocabularies and thesauri, but requires the contextual analysis of keywords. The initial keywords were generated using bootstrapping methods. We have developed a system that uses the Stanford Parser to identify all dependencies in sentences containing one or more keywords. We trained the system using data generated by program officials who performed a thorough retrospective analysis of summary statements from scientific reviews of grant applications. The applications were received over the past several years through several funding mechanisms. We present results that demonstrate the ability of this contextual analysis system to improve the identification accuracy of HRHR research. If successful, the text mining methods can be applied to retrospective identification of unsolicited research proposals to provide a clearer picture of the state of funding for HRHR research.

Introduction

Section 105b of the National Institutes of Health Reform Act of 2006 requires the NIH Director to evaluate the status of HRHR research funded by the NIH and report the results to Congress. In response to this requirement, the High-Risk, High-Reward Implementation Group was chartered to perform a retrospective analysis of grant applications. We have developed a system that identifies HRHR research by analysis of summary statements created during peer review. The process is broken into essentially four steps: extraction of text from key sections of the summary statement; selection of sentences pertaining to HRHR; contextual analysis of sentences; classification of documents as HRHR or not. Text is only extracted from key sections to ensure that the document is classified based on peer-review comments and not on comments from the applicant. Sentences containing keywords that may indicate risk or reward are selected for further contextual analysis. The contexts of keywords found in sentences identify whether the sentence contains the concept of high risk or high reward. Finally, a document is classified by combining the results of all the keywords found. We found our system to approximately 85% accurate.

Methods

The steps required to identify HRHR research are shown in Figure 1. Summary statements are retrieved from the IMPACII system in PDF format. The text is extracted from the PDF file and filtered to select only the "*Resume and Summary of Discussion*" and "*Critiques*" sections. The MetaMap (MMTx) program available from the National Library of Medicine is used to break blocks of text into sentences. The Stanford Parser available from the Stanford Natural Language Processing Group deciphers the grammatical structure of sentences and identifies dependencies between words. A dependency consists of two words and the relationship between the words. As an example, "*very risky*" would have a dependency (very, risky, adverb modification). Only sentences that contains keywords that suggest risk or reward, referred to as primary keywords, are further processed using the Stanford parser. For each sentence processed, a list of dependencies containing primary keywords is generated. The second word in each dependency is compared to a list of secondary keywords that are keyword specific. The secondary keywords specify the context for the primary keyword and indicate whether a sentence implies high-risk (high-reward). The keywords are weighted so that positive values indicate high risk (reward) and negative values indicate low or no risk (reward). Secondary keywords are also weighted, which allows the context to change the meaning of the keyword. Additional levels of keywords can further specify the context. Each keyword is scored by multiplying its weight by the weight of the secondary keywords. For each primary keyword in a summary statement, an individual score is maintained. A document vector, which consists of the keywords as unit vectors with magnitudes equal to the keyword's score, is used to classify the document as HRHR or not HRHR. For this poster, the classifier maps the N-dimensional keyword space into a 2-dimensional risk vs. reward space. Documents with risk and reward greater than a critical threshold are labeled HRHR.

We have preformed additional work into using sentiment analysis to extract sentences that select subjective sentences. The filter uses a naïve Bayes classifier to filter the most subjective sentences. Sentiment analysis is used heavily for analysis of movie reviews to identify if the reviewer liked the movie. The movie reviewer problem is analogous to the HRHR reviewer problem. However, movie reviewers tend to be explicit about their opinions, where peer-reviewer may not overtly comment on risk or reward.

Benefits of the sentiment analysis may include higher accuracy (assuming confounding sentences are removed) and faster processing (which is critical for an analysis of hundreds of thousands of summary statements). Figure 5 shows our results using sentiment analysis. We have a considerable decrease in the number of false positive, however the number of false negatives increase. We plan on tuning the algorithm in an attempt to improve both accuracy and recall.

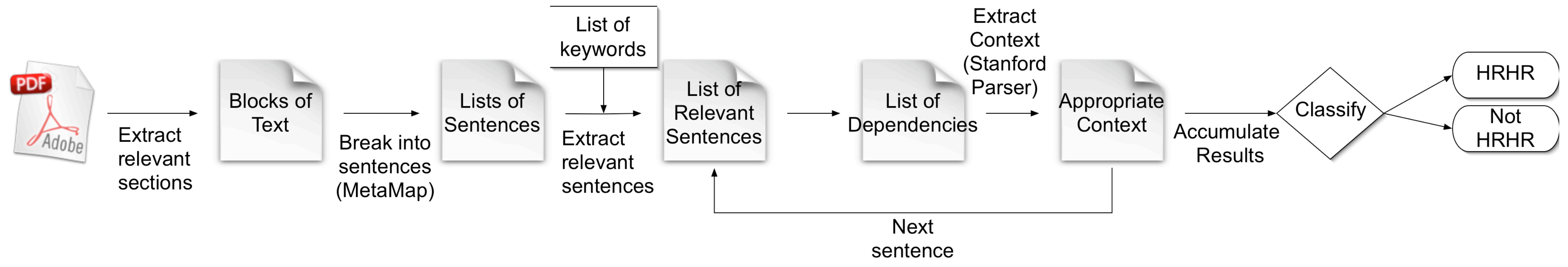


Figure 1. Identification of HRHR summary statements starts with the extraction of the "*Resume and Summary of Discussion*" and "*Critiques*" sections. The text is broken into sentences using Metamap. Sentences contains primary keywords (words denoting risk or reward) are analyzed using the Stanford Parser to produce a grammatical structure and a list of dependencies. If the context is appropriate, we add the keyword to a document vector. The document vector is classified into the HRHR or not HRHR category.

- Finding and development of an alternative and better therapeutic approach for PD will have *significant* impact on this disease
- It is understandable that this approach is of *high* risk
- However, delivering sufficient amounts of genes to the brain to affect the course of the disease is a *major* challenge for neural gene therapy.
- The use of L-DOPA was a breakthrough in the treatment of PD

Keyword	Secondary	score	Reward/Risk
impact	*significant*	+1.0	reward
risk	*high*	+1.0	risk
challenge	*major*	+1.0	risk
breakthrough		+1.0	reward

Keyword Legend
Risk
Reward
Secondary
Negated

Figure 2. The Stanford Parser identifies relationships between words. When keywords are found in a sentence, all relationships containing the the keyword are checked for the presence of secondary keywords. In this example document, four sentences shown above contain primary keywords. The secondary keywords are usually required to mark the context as appropriately high-risk or high-reward. For this summary statement, all keywords were deemed to convey a high-risk or high-reward connotation. The current classification method would identify this project has HRHR.

"In the face of these ambiguities it is not clear that the clinical significance of obtaining mesenchymal progenitor cells from the skin of osteoporotic patients is as great as that suggested by the applicant."

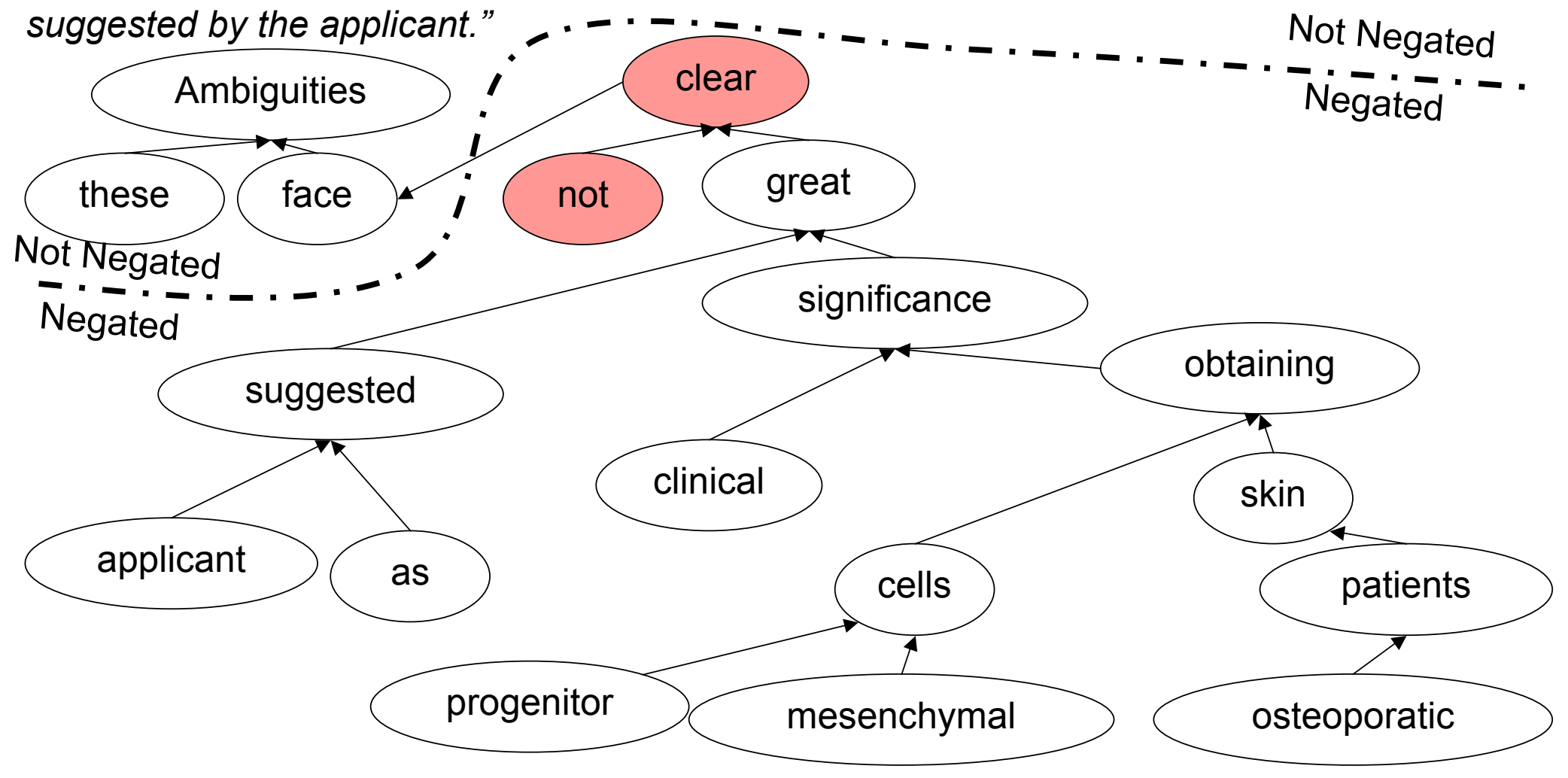
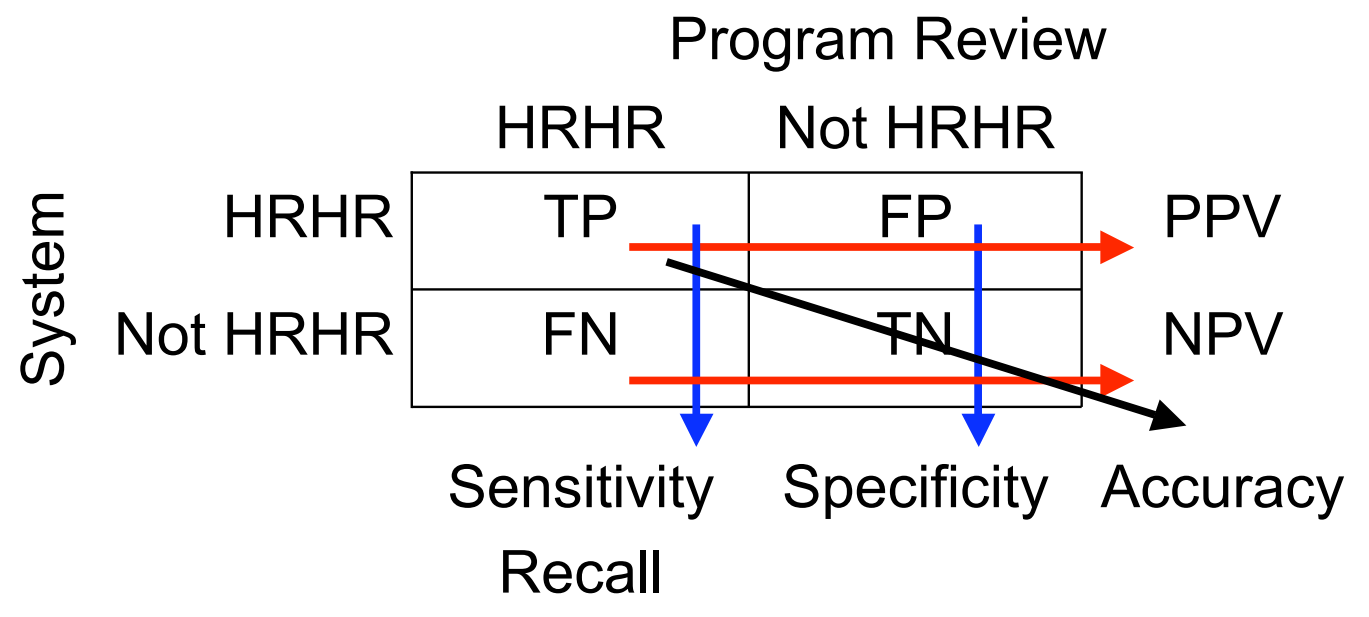


Figure 3. Handling negation is a difficult task for natural language processing. The Stanford parser identifies negation relationships, but does not propagate negation through the sentence. Since the grammar of the sentence is modeled as a tree structure, we use the grammatical structure to identify the relationship that occur under the negated term. Even though the relationship "*progenitor cells*" should be negated, we catch that "*great significance*" is negated. Our algorithm works well at catching negation in the relationships of interest to HRHR classification. In this figure the broken line represents the border between negated and not negated terms.



Program Review

System		HRHR	Not HRHR	Total	
	HRHR	33	49	82	40.24%
	Not HRHR	24	381	405	90.47%
	Total	57	430	487	
		57.89%	88.60%		85.01%

Figure 4. The system results can be measured using different metrics. The table on the left shows a confusion matrix and schematically identifies the meaning of the most commonly used metrics. The middle table shows the a comparison of the system results versus the results of an arduous hand analysis performed by NIH program reviewers. The values for the different metrics are shown on the right table. We focus on two of these metrics, recall and Accuracy. Recall is essentially the probability that an HRHR documents will be identified as HRHR by the system. The accuracy is overall agreement between reviewers and the system. These values can be very different because the documents are overwhelmingly not HRHR. This strong bias makes optimizing the system difficult.

Results

We read program reviewer comments on almost 5,000 summary statements. These comments were used to create an initial list of keywords. Figure 2 shows example output of one summary statement. Only four sentences within the document contain primary keywords. For this document, all keywords indicated high-risk or high-reward. However, both negative keywords and negation is possible. Negation is a difficult task for natural language process, figure 3 show how we handle negation. In our first run, the system analyzed 487 summary statements As seen in figure 4, our results lead to 85% accuracy when compared with program reviewers. We present the most commonly used metrics, however, we believe accuracy and recall are the key metrics. The recall is currently around 58%; however, if we can continue to decrease the number of false negatives, the recall will improve. The other metrics we believe are biased because the number of Not HRHR documents is significantly larger than the number of HRHR documents.

Conclusion

We have produced a system that can identify HRHR projects by their summary statement. The system agrees with reviewers around 85% of the time. The system can be used alone or as a screening service for program reviewer. As a screener, the system can alleviate most of the burden of identifying HRHR documents. Several issues decrease the accuracy of the system. Our system works sentence-by-sentence, when context crosses sentence boundaries, we lose the dependencies (i.e. "*If the application focused on the mechanisms underlying the beneficial impact of social support, this could potentially have a* significant impact*. Unfortunately this is not the case.*"). Also double negation (i.e. "*There is no doubt this is* high-risk"), and grammatical errors (i.e. missing commas) cause trouble that is unlikely to be preventable. However, the results may be slightly better than reported. We assume that program reviewers are the gold standard, but inter-rater agreement may be low.

We have shown the sentiment analysis slightly improves the accuracy of the system results, but additional work is necessary to improve the recall.

Future work

We are running the system for the on selected unsolicited R01 applications going back 5 years. In addition, we are working on statistical learning methods of assessing the quality of the keywords. Changes in the keyword list have dramatic effect on the results.

	Large (n = 487)	
	Semantic Analysis	
	With sentiment extraction	Without sentiment extraction
True Positive	23	33
True Negative	401	381
False Positive	18	49
False Negative	34	24
Recall (sensitivity)	40.35%	57.89%
Accuracy	87.06%	85.01%

Figure 5. Results of additional work using sentiment analysis. A slight improvement of the accuracy is seen at the expense of the recall. The effect is caused by a decrease in the number of false positives at the expense of the number of false negatives.